Consensus modules: modules present across multiple data sets

Peter Langfelder and Steve Horvath

_Eigengene networks for studying the relationships between co-expression modules._

BMC Systems Biology 2007, 1:54
Given several independent data sets, find modules that are present in all (or a specified majority) of the data sets.

Rationale

- Find co-expression patterns common to multiple studied conditions.
- Find common, robustly defined modules across several independent data sets (e.g., from GEO) that study the same conditions.
Finding consensus modules

- Modules group together densely interconnected genes
- Consensus modules group together genes densely connected in all conditions
- Our solution: find the consensus gene-gene similarity and use it with clustering to find modules
Finding consensus modules

- Calibrate input networks to make them comparable
Finding consensus modules

- Calibrate input networks to make them comparable
- For 2-3 sets: Take component-wise minimum

\[
\text{Consensus} = \min_{p\in \text{min}(\)} (\text{Network 1, Network 2})
\]
Finding consensus modules

- Calibrate input networks to make them comparable
- For **4 sets or more**: suitable quantile (for example, quartile)

\[
\text{component-wise quantile} = \text{pquantile()}\]

Consensus

\[
\left( \begin{array}{cccc}
\text{Network 1} & \text{Network 2} & \text{Network 3} & \text{Network 4} \\
\end{array} \right)
\]
Finding consensus modules

Consensus modules are defined from clustering of consensus similarity

\[ = \min (\text{Network 1}, \text{Network 2}) \]
R implementation: blockwiseConsensusModules

- Input:
  - Expression data in "multi-set" format
  - Options for splitting data into smaller blocks if there are too many genes to be handled in one block ("blockwise")
  - Network construction options for constructing individual networks
  - Network calibration options
  - Consensus quantile
R implementation: blockwiseConsensusModules

• Output:
  • Consensus module labels,
  • Gene clustering tree (or trees if the data was split into blocks)
  • Module eigengenes
  • Other diagnostic output

• Introductory tutorial: Consensus analysis of female and male liver expression data (Tutorial II) at labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/Tutorials
Consensus modules and Module preservation statistics

- Consensus modules are by construction present (i.e., preserved) in all (or most) input data sets

- If a module identified in a reference data set is strongly preserved in test data set(s), it would also be a consensus module among the reference and test sets

- Consensus module construction treats all data sets the same; module preservation statistics require a reference and a test data set(s)

- Consensus modules are best suited to answer a different set of questions than module preservation statistics
Application 1: Consensus modules across multiple lung cancer data sets
Eight publicly available lung cancer sets


Eight publicly available lung cancer sets

- Concordance between the sets is poor
- Difficult to find genes consistently related to survival time
- For consistency: restrict analysis to adenocarcinoma
- Since we have 8 sets, we use the quartile instead of the minimum in consensus network construction
Consensus modules across 8 data sets

Red: Upregulated in patients with short survival time
Blue: Downregulated in patients with short survival time
**Association of consensus modules with survival deviance**

<table>
<thead>
<tr>
<th>Module</th>
<th>Meta Z and p</th>
<th>Shedden-MICH</th>
<th>Shedden-HLM</th>
<th>Shedden-DFCI</th>
<th>Shedden-MSKCC</th>
<th>Bild</th>
<th>Tomida</th>
<th>Takeuchi</th>
<th>Roepman</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: defense response</td>
<td>0.34 0.74</td>
<td>0.053 0.26</td>
<td>0.21 0.045</td>
<td>-0.054 0.67</td>
<td>0.15 0.087</td>
<td>0.013 0.46</td>
<td>-0.13 0.88</td>
<td>-0.034 0.62</td>
<td>-0.13 0.82</td>
</tr>
<tr>
<td>2: mitotic cell cycle</td>
<td>5 6.3e-07</td>
<td>0.081 0.17</td>
<td>0.11 0.2</td>
<td>0.21 0.045</td>
<td>0.15 0.085</td>
<td>0.2 0.084</td>
<td>0.37 0.00014</td>
<td>0.36 0.00045</td>
<td>0.11 0.23</td>
</tr>
<tr>
<td>3: extracellular matrix</td>
<td>0.91 0.36</td>
<td>0.0023 0.49</td>
<td>0.18 0.081</td>
<td>-0.059 0.68</td>
<td>0.11 0.14</td>
<td>0.16 0.13</td>
<td>-0.078 0.77</td>
<td>-0.0074 0.53</td>
<td>0.033 0.41</td>
</tr>
<tr>
<td>4: lymphocyte activation</td>
<td>-1.3 0.2</td>
<td>-0.044 0.7</td>
<td>-0.11 0.13</td>
<td>-0.11 0.8</td>
<td>0.0064 0.48</td>
<td>-0.1 0.76</td>
<td>-0.067 0.74</td>
<td>-0.083 0.77</td>
<td>-0.2 0.92</td>
</tr>
<tr>
<td>5</td>
<td>-2.5 0.013</td>
<td>-0.091 0.86</td>
<td>-0.15 0.89</td>
<td>-0.029 0.61</td>
<td>-0.14 0.83</td>
<td>-0.14 0.97</td>
<td>-0.17 0.93</td>
<td>-0.17 0.88</td>
<td></td>
</tr>
</tbody>
</table>

- Cell cycle module shows weak but consistent association with survival time (meta-analysis p-value = 6e-7, highly significant)
- Reflects the known fact that fast-proliferating cancers indicate a poor prognosis for patient
Using consensus modules to study genes related to a clinical trait

- Challenge: how to identify functional categories that associate with a trait (survival time)
- Can use meta-analysis to select genes related to survival time across all data sets, then study their enrichment

![Enrichment in cell cycle genes](image)
Using consensus modules to study genes related to a clinical trait

- Challenge: how to identify functional categories that associate with a trait (survival time)
- Can use meta-analysis to select genes related to survival time across all data sets, then study their enrichment

![Graph showing enrichment in cell cycle genes](image)

- Can also study enrichment of genes with highest connectivity in survival-associated module
Using consensus modules to study genes related to a clinical trait

- Challenge: how to identify functional categories that associate with a trait (survival time)
- Can use meta-analysis to select genes related to survival time across all data sets, then study their enrichment

![Graph showing enrichment in cell cycle genes](image)

- Can also study enrichment of genes with highest connectivity in survival-associated module
- Find much higher enrichment
- Consensus module analysis can lead to better biological insights when pooling several gene expression data sets
Using consensus modules to study genes related to a clinical trait

- Challenge: how to identify functional categories that associate with a trait (survival time)
- Can use meta-analysis to select genes related to survival time across all data sets, then study their enrichment

![Graph showing enrichment in cell cycle genes](chart.png)

- Can also study enrichment of genes with highest connectivity in survival-associated module
- Find much higher enrichment
- Consensus module analysis can lead to better biological insights when pooling several gene expression data sets
Application 2:
Consensus modules across 4 brain regions in Huntington's Disease patients and controls
Data

- Huntington's disease primarily affects motor skills
- Biggest changes are observed in Caudate Nucleus (CN), much smaller changes in Cortex and Cerebellum (CB)
- Disease causes dying of neurons and increase in astrocytes/oligondendrocytes (inflammatory response)
- Here: consensus analysis of the 4 data sets

Consensus network and modules

- Two large branches that correspond to neurons and astro/oligodendrocytes

Consensus clustering and gene significance

- Red: underexpressed in HD; blue: overexpressed in HD
Significance of modules for disease status

### Consensus module significance for disease status

<table>
<thead>
<tr>
<th>Module</th>
<th>Caudate Nucleus</th>
<th>Motor Cortex</th>
<th>Association Cortex</th>
<th>Cerebellum</th>
</tr>
</thead>
<tbody>
<tr>
<td>3: red_M19__CTX</td>
<td>-0.11 0.4</td>
<td>0.12 0.5</td>
<td>0.02 0.9</td>
<td>0.11 0.4</td>
</tr>
<tr>
<td>10: brown_M3_Astrocytes__HumanMeta</td>
<td>0.02 0.02</td>
<td>0.46 0.01</td>
<td>-0.07 0.7</td>
<td>0.18 0.2</td>
</tr>
<tr>
<td>7: blue_M2_Oligodendrocytes__HumanMeta</td>
<td>0.29 0.02</td>
<td>-0.16 0.4</td>
<td>-0.15 0.5</td>
<td>-0.1 0.5</td>
</tr>
<tr>
<td>105968.REACTOME_Formation_of_a_pool_of_free_40S_subunits_Re</td>
<td>0.55 2e-06</td>
<td>0.51 0.004</td>
<td>0.28 0.2</td>
<td>0.02 0.9</td>
</tr>
<tr>
<td>2: brown_M3_Astrocytes__HumanMeta</td>
<td>0.69 1e-10</td>
<td>0.59 6e-04</td>
<td>0.2 0.3</td>
<td>0.38 0.004</td>
</tr>
<tr>
<td>4: pink_M10_Microglia(Type1)__HumanMeta</td>
<td>0.48 5e-05</td>
<td>0.25 0.2</td>
<td>0.13 0.5</td>
<td>0.29 0.03</td>
</tr>
<tr>
<td>1: red_M11_Neuron__HumanMeta</td>
<td>-0.66 2e-09</td>
<td>-0.51 0.004</td>
<td>-0.18 0.4</td>
<td>-0.45 5e-04</td>
</tr>
<tr>
<td>9: membrane-bounded organelle</td>
<td>-0.03 0.8</td>
<td>-0.2 0.3</td>
<td>-0.12 0.6</td>
<td>-0.25 0.06</td>
</tr>
<tr>
<td>5: green_M10_GlutamatergicSynapticFunction__CTX</td>
<td>-0.6 1e-07</td>
<td>-0.45 0.01</td>
<td>-0.14 0.5</td>
<td>-0.22 0.1</td>
</tr>
<tr>
<td>6: DownWithAlzheimers_Ballock__ADvsCT_inCA1</td>
<td>-0.58 3e-07</td>
<td>0 1</td>
<td>0.18 0.4</td>
<td>0.26 0.05</td>
</tr>
</tbody>
</table>

**Message:** The disease effect is strongest in Caudate Nucleus and Motor Cortex, consistent with HD manifestations.
Significance of modules for disease status

The modules relate to disease status similarly across all tissues.
Study meta-networks built from eigengenes of consensus modules

- Recall: modules are represented by their eigengenes (singular vectors obtained from SVD)
- Each consensus module has one eigengene in each data set
- In each data set: correlation among module eigengenes gives a bird-eye view of the entire gene network: a correlation matrix of 12k genes is reduced to a correlation matrix of 10 eigengenes
- Correlation of eigengenes reflects how the underlying pathways, processes, cell types, etc work together
- It may be interesting to study how eigengene correlation changes between data sets

Langfelder and Horvath, *Eigengene networks for studying the relationships between co-expression modules*. BMC Systems Biology 2007, 1:54
Preservation of eigengene networks between brain regions

- Eigengene network defined as a signed network with power $\beta=1$:
  \[ A_{ij} = \frac{1 + \text{cor}(E_i, E_j)}{2} \]

- Preservation network: measures how much eigengene correlation varies among data sets
  \[ \text{Pres}_{ij}^{(1,2,\ldots)} = 1 - \left[ \max \left( A_{ij}^{(1)}, A_{ij}^{(2)}, \ldots \right) - \min \left( A_{ij}^{(1)}, A_{ij}^{(2)}, \ldots \right) \right] \]

- Mean preservation: measures overall preservation of eigengene networks among data sets
  \[ D^{(1,2,\ldots)} = \text{mean}_{i<j} P_{ij}^{(1,2,\ldots)} \]
Preservation of eigengene networks between CN and Association CTX

- Preservation is relatively high: D=0.88
- Preservation of networks among neuronal modules is lower than the networks among inflammation-related modules
- We did not discover a cure for HD, but certainly have "Food for thought" for biologists
What have we learned?

- Cancer application: consensus analysis identifies a module consistently related to survival time

- The module provides a cleaner biological interpretation than genes identified using standard meta-analysis

- Huntington's disease application: Consensus module analysis shows that HD affects different brain regions in a broadly similar manner but also shows differences in the way the regions are affected

- Consensus eigengene network analysis provides a way to study commonalities and differences in network organization of gene expression or other genomic data
• Consensus modules and eigengene networks:

  *Eigengene networks for studying the relationships between co-expression modules.*
  BMC Systems Biology 1:54
  labs.genetics.ucla.edu/horvath/htdocs/CoexpressionNetwork/EigengeneNetwork/

• Cancer application of consensus modules:

  *When Is Hub Gene Selection Better than Standard Meta-Analysis?*
  labs.genetics.ucla.edu/horvath/CoexpressionNetwork/MetaAnalysis/